THE APPLICATION OF MULTIPLE FRAME METHODOLOGY *

Introduction

A Multiple frame sample survey is one relying upon the joint use of two or more sample frames. The Statistical Reporting Service (SRS) in the U.S Department of Agriculture uses multiple frame sampling methodology on many surveys of major importance. For example, surveys to provide estimates of livestock inventories, farm labor and wage rates, crop acreages, as well as expenditure surveys are based on the methodology of multiple frame sampling.

This paper will discuss reasons for using multiple frame sampling designs, problems involved in the application of the methodology, estimators used, and research underway in the Department of Agriculture to improve the procedures.

Before discussing the application of multiple frame sampling, it is probably appropriate to review its development in the USDA. One of the first SRS multiple frame surveys was a poultry marketing survey conducted by the Maryland Statistical Office in 1956.[1] Like other multiple frame surveys conducted during this time period, samples were selected from both frames, but whenever a unit sampled from the area frame was also found to be in the list frame, it was discarded. In the early 1960's SRS entered into a cooperative agreement with H.O. Hartley at Iowa State University for the original development of theory for multiple frame sampling. This led to Dr. Hartley's paper entitled "Multiple Frame Surveys," which was presented to the Minneapolis meetings of the American Statistical Association in 1962.[6] The significant result of this paper was a weighting procedure to combine information collected from the overlapping frames, meaning it would not be necessary to discard any information to estimate the population total. Subsequent agreements with Iowa State and Texas A & M Universities resulted in an extension of this basic theory. Results of much of this work are presented in Robert S. Cochran's Ph.D. thesis, "Theory and Applications of Multiple Frame Surveys."[2] Since then, others have studied design and estimation problems associated with multiple frame surveys. A paper by Fuller and Burmeister provides an excellent reference for most of the theoretical work completed to date.[4] Most of the work attempts to improve the estimator provided by Hartley.

Multiple frame sampling as used by SRS involves the joint use of two frames, an area frame and a list frame. The area frame is the complete frame or the 100 per cent frame. Every farm operator via a sampling unit or segment of land has a chance to be selected from this frame. The frame is generally stratified by land use, beyond that the only efficient means to improve sampling precision is to increase the sample size. This frame is usually the more expensive to use for obtaining survey data, because the survey must usually be completed by personal enumeration.

The list frame is a list of names of farm operators. It may also contain information by which to stratify and therefore improve the sampling efficiency. In many survey situations, the use of cheaper mail and telephone data collection procedures can be used. However, this frame is incomplete and will not provide information for the entire population.

* By Fred A. Vogel, Statistical Reporting Service, USDA, 1974.

Therefore, the use of multiple frame sampling is applicable to this situation. It allows the maximum use of the list frame, which if constructed properly will be considerably more efficient than the area frame, yet when in combination with the complete area frame the ability to obtain the most efficient estimates for the population of interest.

However, in a recent paper,[7] Hartley stated that "Perhaps the most important reason why multiple frame surveys have not been used more extensively in the past are the operational problems in their implementation." Hartley also discusses some of the problems that occur. The main purpose of this paper is to summarize in more detail some of the problems and a comparison of different methods used to handle them. A more detailed discussion of these problems will be found in a paper entitled "Surveys With Overlapping Frames- Problems in Application." [9]

Problems in Application

Mutliple frame surveys are subject to all operational problems that plague single frame surveys, however by their very design, problems unique to multiple frame surveys also occur. These problems arise from the basic assumptions involved in multiple frame sampling:

a. Every element of the survey population must be included in at least one of the frames.

b. It must be possible to determine for every selected sample unit whether or not it belongs to any other sample frame, i.e., the overlap between frames must be determined.

The first assumption is satisfied by using the area sampling frame. The latter assumption leads to one of the most critical aspects of a multiple frame survey. Sometime during the survey process it is necessary to determine for every sampled unit whether it could have been selected from another frame also being used. Since the area frame is conceptually a complete frame, the overlap between the two frames is identified by determining whether every farm operation found in the area frame sample could also have been selected from the list frame. Prior to the survey it is not known which area frame units are on the list and those which are not on the list. Therefore, during the survey it is necessary to obtain sufficient information about every farm operator found in the area frame sample so that it can be linked with a list unit if it is in the list.

The available theory does not tell us how this overlap determination is to be made, it only gives us alternative estimators to use once the determination is made.

The need to determine the overlap between the sample frames affects many of the survey procedures. For example, the list frame contains names of potential farm operators. While the sampling unit is a name, the reporting unit is all land operated by the particular name. The data collection procedures must determine the association between a name and a unit of land. The area frame sample units are small areas of land called segments. Each sample segment is screened for farm operations. A sample segment may contain portions of 3-5 farming operations. The names of the farm operators associated with each parcel of land or operation found inside the segment are obtained

during the survey. Here, since the sample unit is land, it is necessary to associate a name with each unit of land.

In practice, it must be assumed that an area of land can be represented by a name. Then, in the multiple frame context, the overlap of land areas represented by both sample frames is identified by matching names associated with the land. This is probably the most difficult factor involved in a multiple frame survey. Errors in this determination are not considered in the estimation phase; thus, they fall into the area of nonsampling errors. The name matching operation can be completed manually or by a computer method or record matching as described by Fellegi and Sunter. 3/ Which ever procedure is used requires certain decision logic about what is a match and what is a nonmatch.

Two terms to be used are now defined. The area frame sample (the 100 percent frame) must be divided into two domains for multiple frame estimation:

a. Nonoverlap Domain - This domain consists of population units or farms found via the area frame sample that are not in the list frame.

b. Overlap Domain - This domain contains sample units that are also in the list frame. These farm operations in the area frame sample also had a chance to be selected from the list frame. To see why the domain determination is required it is necessary to look at the estimators.

An unbiased estimator using the area frame is: $_a\hat{X} = \sum_h \frac{aN_h}{an_h} \; a\acute{X}_h$ where $\frac{aN_h}{an_h}$ is the reciprocal of the probability of selecting a sample unit in the area frame and $\acute{X}_h$ is the sample total for a particular stratum. The area frame estimator can also be written as:

$$_a\hat{X} = \sum_h \frac{aN_h}{an_h} \; (a1\acute{X}_h) = a1\,\hat{X} + a2\hat{X}.$$

Here $a_1\hat{X}$ is an estimate of the incompleteness of the list frame or the nonoverlap domain of the area frame. The $_{a2}\hat{X}$ is the area frame estimate of the population also represented by the list frame (overlap domain).

A multiple frame estimator suggested by Hartley for the case where one of the frames is complete, is, $\hat{X} = {}_{a1}\hat{X} + P\,{}_{a2}\hat{X} + Q_b\hat{X}$ where $_b\hat{X}$ is an estimate of the overlap domain based on the list frame sample and the weights P and Q are such that P + Q = 1.

A simplier multiple frame estimator is one where P = 0 and Q = 1. Then, no information from the area overlap domain is utilized. However, in either case, it is necessary to divide the area frame into the two domains. Next, some examples illustrating problems that are encountered, different alternatives for defining the domains, and the consideration of the problems in the estimators will be discussed.

There are two factors contributing to the problems with domain determination or determining whether a farm operation found in the area frame is also in the list frame.

a. One relates to the matter of duplication in the list. Several procedures have been devised for using computers to remove the duplication; however, the problem will continue to exist. The survey procedures for identifying and adjusting survey data when duplication exists in the list frame must be considered in a multiple frame survey, not only for the estimation but also for domain determination.

b. Some larger farming operations, such as partnerships or corporations contain several individuals that may report for the entire operation. These individuals can appear on the list frame either singularly or in combination with other names. This poses a problem in estimation for the list frame. It also poses a problem in determining whether a given operation is overlap with the list frame or not.

The following table illustrates some of the problems encountered when identifying the overlap between the two frames by matching names.

Table 1--Examples that occur when determining the overlap between the area and list sample frames.

Name(s) associated with land in area frame segment	Name(s) in list frame that may represent land in area segment
(1) James Smith	James E. Smith
(2) James Smith	James E. Smith Jim Smith
(3) Bill, Bob, Joe, and Sam Jones	Sam Jones Robert Jones

The first example involves a unit of land operated by only one individual. A decision is required to determine whether the land operated by James Smith as reported in the area frame sample overlaps land operated by James E. Smith in the list frame. They may be the same and the interviewer forgot to ask for a middle initial during the interview. This problem can be resolved by reviewing and comparing the complete address and/or reported data in the area frame vs. control data in the list. It may also involve a reinterview. The second example still involves an individual operation, but also involves potential duplication in the list. The same factors considered above need to be considered. However, this problem also involves the estimator which will be discussed shortly.

The last example involves an operation jointly operated by several people These operations present some of the more perplexing problems which involve both survey procedures and estimators. There are four names associated with a parcel of land in an area frame sample segment. Two of the four names appear in the list frame. Does the parcel of land in the area

frame overlap with land operated by Sam Jones, or with Robert Jones, or with both? Can the land be reported twice from the list frame? Not only do we have the problem of determining overlap between the two frames - there is also the possibility of duplication in the list. In an operational survey, rules must be established so that such problems as above are handled consistently. Three alternate rules are compared below and will be applied to the third example shown in the table.

Rule A

This rule is based on the following assumptions:

a. Each partner will report for the entire operation and correctly identify all of his partners.

b. If more than one partner appears somewhere in the list frame, he will be identified.

c. Duplication in the list will be identified.

Since we assume that each partner will report for the entire operation, the parcel of land found in the area frame overlaps the operation represented by the two names on the list. However, there remains the problem of duplication within the list.

Different procedures are available for handling this duplication in the estimation One is presented by Gurney and Gonzalez [5] where the number of times a given operation is duplicated is not known. Another method has been developed by Rao [8] for the case where the number of times an operation can be selected from the frame is known.

It will be assumed we can determine the number of times every selected unit could have been sampled. This is done by matching each name in the list sample with the remaining names in the list frame. Controls are also built into the survey questionnaire to aid in the detection of possible duplication. For example, each respondent is asked whether he is known by any other name or if any other names are associated with his operation. Rao's procedure was developed for the case where there is no stratification in the frame. His estimator for the list frame would be:

$_b\hat{X} = \sum_{b}^{n} \frac{{}_bN}{{}_bn} \frac{{}_bX_i}{{}_bA_i}$ where ${}_bN$ and ${}_bn$ are the total number of names and number of selected names respectively from the list frame. ${}_bA_i$ is the total number of times a given unit (farm operation) can be selected from the frame. In this example ${}_bA_i = 2$.

The procedure outlined by Rao can be extended to the case where the duplicated names in the list frame are in different strata.

Again, we wish to estimate the population total (X) for the list frame from a sample. The population value can be obtained by summing over the population as follows:

$_bX = \sum_{h} \sum_{i}^{{}_bN_h} \frac{{}_bX_{hi}}{{}_bA_{hi}}$. Here, ${}_bA_{hi}$ is the total number of times a ${}_bX_{hi}$ unit can be

selected from the list frame. It is assumed the b^A_{hi} factor can be determined correctly from the sample. The duplicated operation is included in the tabulation every time it is selected.

The estimator $\sum_h \sum_i^{a^n_h} \frac{b^N_h}{b^n_h} \frac{b^X_{hi}}{b^A_{hi}}$ for the case where the list duplication occurs in different strata can be shown to be unbiased by writing $\sum_h^{b^N_h} \frac{b^N_h}{b^n_h} \frac{b^X_{hi}}{b^A_{hi}} \cdot t_{hi}$ where $t_{hi} = 1$ if the $_i$th unit is selected, 0 otherwise.

Then $E(\hat{X}) = X$ because the sample is selected independently within strata and each duplicated operation is given the value X_{hi} no matter how many times it is selected.

The result is simply a rule for assigning a portion of the duplicated operation to each stratum from which it could have been selected. Here $b^A_{hi} = 2$ is the total number of times the operation could have been selected from the entire list frame. The multiple frame estimator is then obtained by adding the list estimator $_b\hat{X}$ to the area frame portion, i.e.

$X = {}_{a1}\hat{X} + P\ {}_{a2}\hat{X} + Q\ {}_b\hat{X}$. The success of this estimator depends on the ability to correctly define the domains in the area frame. If the assumption that each individual will report for the entire partnership does not hold in practice, the estimator becomes biased. This occurs because $_{a2}\hat{X}$ will be estimating for an operation that is not represented by the list. It also means that the b^A_{hi} weights are incorrect.

A second rule follows.

Rule B

This rule relies on fewer assumptions than Rule A. It only requires that each individual partner report for the entire operation and correctly identify all his partners. Operational procedures differ however and are illustrated:

a. The total number of partners associated with the parcel of land in the area frame sample unit are identified. The number is designated by a^A_{hi}.

b. The number of partners associated with the area frame sample unit and that are also on the list frame is determined. This number is b^A_{hi} as defined for procedure A.

c. A weighting factor is determined for assigning a portion of the operation to the list frame and a portion to the area frame. The factor to be applied to the area frame for this unit is $1 - \frac{b^A_{hi}}{a^A_{hi}} = 1 - \frac{2}{4}$

The factor applied to each of the duplicated list frame sample units is then $1/a^A_{hi} = \frac{1}{4}$. The multiple frame estimator then becomes

$$\hat{X} = \Sigma^{a}\Sigma^{n}_{h} \frac{a^{N}_{h}}{a^{n}_{h}} \left(1 - \frac{b^{A}_{hi}}{a^{A}_{hi}} \right) a^{X}_{hi} + \Sigma^{b}\Sigma^{n}_{h} \frac{b^{N}_{h}}{b^{n}_{h}} \frac{b^{X}_{hi}}{a^{A}_{hi}}$$

Note that if $b^{A}_{hi} / a^{A}_{hi} = 1$ or if $b^{A}_{hi} = 0$ for every sample unit the multiple frame estimator is $\hat{X} = {}_{a1}\hat{X} + {}_{b}\hat{X}$ which is the result when the P + Q weights are 0 and 1, respectively. The two rules presented so far both provide unbiased estimators. The main difference in the rules is in the complexity of their application. Although all of the procedures result in unbiased estimators, the important point is that they may differ in the bias resulting from the breakdown of the assumptions.

The following rule relies upon a different set of assumptions for defining overlap between the area and list frames.

Rule C

We are still referring to problem 3 as illustrated in Table 1. The assumptions here are:

a. An individual name or the name of a single person on the list represents a unique land operation only associated with that name. More specifically, the name Sam Jones can only represent land operated soley by Sam Jones. It cannot represent land operated jointly by himself and others.

b. If the individual name does not have a unique operation it is considered to be out of business.

When applying these assumptions to problem 3, we obtain the following results:

a. The parcel of land in the area frame sample operated by the four people mentioned does not overlap with a list frame unit. The operation would be overlap only if a list unit consisted of the four names.

b. There is no duplication in the list frame since each name will only report for land unique to itself. Thus, the estimator does not rely on the b^{A}_{hi} factor.

Rule C does not rely upon the assumption that every person in a partnership operation will report for the entire operation. Instead, it relies on the assumption that an individual name will only report for individual data. As a result, the amount of overlap between the list and area frames is decreased by Rule C which then should increase the size of the nonoverlap domain. This is especially true as the list frame becomes more and more out of date - meaning that changes in names of operations or changes in partners will result in fewer matches between the two frames.

Results

As was stated before, multiple frame estimation requires that the overlap between the sample frames be identified. In other words, the components $a1\hat{X}$ and $a2\hat{X}$ must be accurately determined for the multiple frame estimator to be valid.

The variance estimator

$$VAR\ \hat{X} = Var_{a1}\ \hat{X} + P^2 Var_b \hat{X} + 2PCOV\ a1\hat{X}a2\hat{X}$$

only measures the variability due to random sampling. It gives no measure of the accuracy of the overlap determination. The inaccuracy of the overlap determination falls in the realm of nonsampling errors which are difficult to measure. Since many of the problems associated with overlap determination also affect procedures for handling duplication in the list, additional nonsampling errors can occur.

The rules illustrated above were used in an operational multiple frame survey designed to estimate total hogs and pigs on farms. The purpose was to examine difficulties involved when applying each procedure and to measure the differences in the estimates and sampling errors resulting from each rule.

The sample for the survey consisted of about 2,200 farming operations from the area frame sample. Names associated with the 2,200 farming operations were matched with names on a list containing some 80,000+ potential farm operators. Area frame operations with names matching list frame names constituted the overlap domain. The domain determination was done using each of the three procedures.

A sample of 1,600 names was selected from the list frame and also included in the survey. Partnership operations and duplication in the list were processed using each of the three procedures. Then a multiple frame estimate based on each procedure was computed. The results appear in the following table.

Table 2--Multiple Frame Estimates and Sampling Errors resulting from three Rules for defining overlap between sample frames

Rule	Multiple Frame estimates	Sampling Error
A	13.2	.5
B	13.4	.5
C	14.1	.6

Rules A and B gave similar results, but then their basic assumptions are about the same. Rule C differed considerably in the results. Remember that the three rules were all applied to the same sample and that unbiased estimators were used. The sampling error of the difference between any two of the estimates was about .2. This shows that Rule C resulted in a significantly different estimate from that resulting from Rules A and B.

The larger estimate resulting from Rule C resulted primarily from an increase in the estimate from the nonoverlap domain. Theoretically, any increase in the nonoverlap domain should be offset by a decrease in the overlap domain and list frame estimate; however, this did not occur. This indicates a problem with a key assumption

Rule A & B: Every individual in a partnership will report for the entire partnership and will correctly identify all other partners.

Rule C: An individual will only report for individual operations.

The procedures can only be compared by evaluating the total error involved, i.e., sampling error plus nonsampling error. The problem is that all three procedures involve some subjectivity. This involves the accuracy with which the respondent can define his operation whether it be an individual or a partnership operation and can be affected considerably by the questionnaire design.

Since Rules A & B resulted in the lower estimates, the assumption that every individual in the partnership will report for the entire operation may not be met. Rules A & B also involve more subjectivity and complexity because of the necessity of determining A_{hi} factors for partners and for duplication. Rule C is less complex and therefore should be easier to implement in an operational survey. However, the assumption for Rule C may also be failing; that is, an individual may report for more than his individual operation.

Research Activities

The primary research activity presently underway involves the different rules (A, B, and C) and their underlying assumptions as they are used to determine the overlap between the sampling frames. The operational multiple frame surveys currently being conducted by SRS rely on either Rules A or B. The basic assumptions for both are very similar; thus they differ primarily in the execution. Therefore, the primary research effort underway is to contrast Rule C with A & B. The purpose of this research is two fold:

a) To determine which procedure (A, B, or C) can be implemented with the least amount of nonsampling errors
b) To improve the questionnaire design to make the proper association between the reporting unit and the sampling unit for the procedure to be used.

A brief description of this research activity follows:

a) Operational surveys are conducted using a "current" questionnaire and assumptions based on Rule A or B. This acts as the control. This questionnaire also contains sufficient data to apply the assumptions based on Rule C.

b) Independent samples will be selected and surveyed using a "Test" questionnaire and the assumptions based on Rule C.

c) After the initial survey, subsamples will be selected from both the operational and "test" samples. Both subsamples will be reinterviewed to determine if the assumptions were met.

The analysis of experiments such as these will answer questions about the appropriateness of the assumptions and the questionnaire design. The use of independent samples and reinterviewing will be an on going program to evaluate survey procedures and to provide a quality check.

The SRS research effort also involves the development of more efficient multiple frame estimators. The estimator currently used is essentially that developed by Hartley in 1962 i.e.

$$\hat{X} = {}_{a1}\hat{X} + P_{a2}\hat{X} + Q\ {}_{b}\hat{X}$$

Often times the weights are such that p = 0 and Q = 1 which reduces the estimator to

$\hat{X} = {}_{a1}\hat{X} + {}_{b}\hat{X}$. This is based on considerable analysis which has indicated the optimum value of P is near zero.

However, the primary emphasis in the research concerning estimators involves a variation of the Hartley and the Fuller-Burmeister estimators that utilizes the stratification in the list frame. The list frame estimator ${}_{b}\hat{X}$ is the result obtained from a stratified sampling design. The overlap domain from the area frame can be further divided into as many sub domains as there are strata in the list frame.

Then one possible estimator is

$\hat{X} = {}_{a1}\hat{X} + \sum_{h}^{k} (P_h\ {}_{a2}\hat{X}_h + Q_h\ {}_{b}\hat{X}_h)$. Preliminary analysis has shown that the optimum P_h and Q_h weights differ considerably between strata and also differently from the overall P & Q values. The preliminary analysis also indicates this estimator may show considerable gains over that based on an overall P & Q weighting scheme. An interesting consequence is that the optimum P value may be near 1.0 for the small size group strata. Therefore, this analysis will be accompanied by an in depth cost analysis and development of cost functions to determine the optimum P_h and Q_h as well as the optimum sample allocation between the frames.

last, but certainly not least is the research effort to develop record linkage procedures to handle the problems of within list duplication as well as the name match procedure to identify the overlap between sample frames. This includes the development of models to estimate the likelihood of matches vs. nonmatches. Hopefully then, these likelihoods can be implemented into the estimators and their variances.

REFERENCES

1/ Caudill, Charles E., "Concepts of Multiple Frame Sampling" proceedings of the National Conference of the Statistical Reporting Service, U.S. Department of Agriculture, April 1970.

2/ Cochran, Robert S. "Theory and Applications of Multiple Frame Surveys," Ph.D. Thesis, Iowa State University, 1965.

3/ Fellegi, I.P. and Sunter, A.B. (1969), "A Theory for Record Linkage," Journal American Statistical Association 64, pp. 1183-1210.

4/ Fuller, Wayne A. and Burmeister, Leon F., "Estimators for Samples Selected From Two Overlapping Frames," proceedings of the Social Science Section of the Montreal Meetings of the American Statistical Association, 1972.

5/ Gurney, Margaret and Gonzalez, Maria Elena, "Estimates for Samples From Frames Where Some Units Have Multiple Listings." Proceedings of the Montreal Meetings of the American Statistical Association, 1972.

6/ Hartley, H.O., "Multiple Frame Surveys," paper given at Minneapolis meetings of the American Statistical Association, September 1962.

7/ Hartley, H.O., "Multiple Frame Methodology and Selected Applications," Sankhyā, the Indian Journal of Statistics, 1974 Volume 36.

8/ Rao, J.N.K., "Some Non-Response Sampling Theory When the Frame Contains an Unknown Amount of Duplication," Journal of the American Statistical Association, March 1968 (87-90).

9/ Vogel, Frederic A., "Surveys With Overlapping Frames-Problems in Application," Presented at the 135th Annual Meeting of the American Statistical Association Atlanta, Georgia, August 25-28, 1975.